IOP Conf. Series: Journal of Physics: Conf. Series **913** (2017) 012003 doi:10.1088/1742-6596/913/1/012003

Big Data Analytics and Pattern Recognition Methods in the Problem of Optimization of Technological Processes in Metallurgical Production

D Gainanov[1] and D Berenov[2]

[1] The head of department "Big Data Analytics and Methods of Video Analysis", Ural Federal University, Ekaterinburg, Russia
[2] PhD student, Ural Federal University, Ekaterinburg, Russia

E-mail: damir.gainanov@gmail.com

Abstract. The paper considers the problem of predicting the quality of metallurgical production in a general formulation, when a huge amount of historical data of technological processes and product quality can be partitioned into a set of discrete classes of efficiency. Assuming that in the production process the technological process route can be changed, the problem of choosing the optimal route becomes important. When constructing the decision tree that recognizes the class of efficiency of the technological process, a special criterion for optimality of the partitioning of the set of classes of efficiency into two classes is introduced corresponding to the left and right branches of the decision tree in the node under consideration. The partitioning obtained by the proposed approach is close to optimal and can form the basis for constructing a decision tree, with the help of which a route is chosen to continue processing.

1. Introduction

The methods of Big Data analytics are widely used in various fields including the field of production optimization. Metallurgical production, due to a number of technical features, requires special attention to the quality of products. The problem of increasing the efficiency of production in terms of product quality can be investigated using a mathematical methods, since many routing processes can be described in terms of graph theory. At the same time, an adequate solution to the problem of predicting the quality of metallurgical production should take into account the historical experience of the production. Thus, the research of the problem covers not only the field of development and formalization of the mathematical model of the production process, but also the field of development of methods for analyzing and processing of Big Data.

In work [1] the problem of predicting the defects in metallurgical production was formalized and effective decision algorithms were obtained. This problem is a special case of the overall quality prediction problem and is investigated in this work using Big Data technologies. To solve this problem, methods such as an alternative covers or committees constructions, described, for example, in [2–6], can be used in the formulation of the research of infeasible systems. In this paper, we research the problem of predicting the quality of metallurgical production in a general formulation, which reduces to the problem of pattern recognition in a geometric formulation. The results in the solution of the problem of pattern recognition in a geometric formulation are given,

for example, in [7, 8]. In the present paper the formal justification and the mathematical formulation are given. Data on completed technological processes form a training sample of the supervised learning. In view of the large dimensionality of the training sample, the need for developing effective methods for analyzing Big Data follows in an obvious way. A special criterion is introduced for the optimal partition of the set of classes of efficiency of technological processes, which makes it possible to significantly reduce the computational costs of constructing a decision tree. A formal description of the algorithm and the results of the implementation on the test case are given.

2. Basic definitions

In the process of metallurgical production, a certain unit of production (PU) is processed sequentially on a number of technological aggregates. The methods of graph theory provide an adequate tool for modelling production processes, including metallurgical ones.

Definition 2.1 A directed graph $\vec{G} = (\mathrm{A}, E)$, the set of vertices of which corresponds to a set of technological aggregates, is called an infrastructure graph, if $(A_1, A_2) \in E$ if and only if the output PU of the aggregate A_1 can serve as the input PU for the aggregate A_2.

The sequence of the aggregates forms a technological route (TR) for processing the PU. At the same time, according to the results of the processing, for each unit of the PU, values are assigned for a given set of parameters. In these terms, TR can be defined as a path in an infrastructural graph, that is, $P = (A_{i_1}, A_{i_2}, \ldots, A_{i_k})$. We denote the set of all TRs by $\mathrm{P} = \{P_1, \ldots, P_k\}$.

Definition 2.2 Sequence

$$AI_i = \left(A_{i1}, \mathrm{P}_{i1}(AI_i), \ldots, A_{is}, \mathrm{P}_{is}(AI_i)\right),$$

where $\mathrm{P}_{ij}(AI_i)$ is a set of parameter values for a PU in a specific implementation of TR AI_i, is called a completed technological route (CTR).

A set of CTR is formed continuously in the process of production activity. Thus, the problem arises of processing historical data. In this case, in view of the fact that each PU is characterized by a set of parameters of large dimension, the problem arises of processing of Big Data. In this paper we propose an approach to solving the problem of optimization of technological processes using methods of graph theory and pattern recognition.

We denote by $G^k(A)$ the set of all vertices A of the graph $\vec{G}$ such that there exists a simple path from the vertex A to the vertex A' of the length $k-1$. The geometric meaning of this notation is close to the notion of a neighborhood of the k-th order in the graph.

Definition 2.3 Technological pyramid $\mathrm{Pir}(\vec{G}, A)$ with a root vertex A is a subgraph of a graph $\vec{G}$, generated by a set of vertices $\langle A \cup G(A) \cup G^2(A) \cup \ldots \cup G^k(A) \rangle_{\vec{G}}$ such that any directed path P of the graph $\vec{G}$, starting at the vertex A, lies entirely in this subgraph.

Definition 2.4 A fork-vertex of a graph (subgraph) is a vertex $A \in \mathrm{A}$, such that $|G(A)| > 1$.

Definition 2.5 A terminal vertex of a graph (subgraph) is a vertex, from which no arc leaves in this graph (subgraph).

Let A' be the set of terminal vertices of the technological pyramid $\mathrm{Pir}(\vec{G}, A)$. For each terminal vertex $A_i \in \mathrm{A}'$ there is some PU, which is the output PU for this vertex, and there may be several EPs depending on the type of CTR, as a result of which the given PU has been obtained.

Definition 2.6 An CTR is called a productive if the PU at the output of the terminal vertex $term\left(AI_i\right)$ of the CTR AI_i is one of the types of the final product deliverable to the market.

The final PU obtained as a result of passing the productive CTR is characterized by a quality indicator. For example, defective or suitable parts determine the quality of the final PU.

Definition 2.7 The effectiveness of a productive CTR AI_i is the value calculated as

$$Ef\left(AI_j\right)=\frac{\mathrm{Price}\left(term\left(AI_j\right)\right)-\mathrm{C}\left(term\left(AI_j\right)\right)}{\mathrm{C}\left(term\left(AI_j\right)\right)}$$

where $\mathrm{Price}\left(term\left(AI_j\right)\right)$ characterizes the market price of the unit of measurement of the final PU and $\mathrm{C}\left(term\left(AI_j\right)\right)$ is cost of production of the unit of measurement of the final PU.

The investigated process of metallurgical production allows the choice of TR for the continuation of processing of PU. For different productive CTRs, the sequences of technological aggregates may coincide, however, the PU parameters are different for different productive CTRs, which entails a difference in efficiency indicators. Thus, the practical meaning of the problem is to choose such a TR to continue processing the PU, so that the final efficiency indicator (that is, the quality index of the final PU) will reach its optimal value. The basis of this choice, obviously, can be the historical experience accumulated in the production process.

3. The problem of forecasting the quality of production

Consider the fragment of the infrastructure graph $\vec{G}=\left(\mathrm{A},E\right)$ presented in Figure 1.

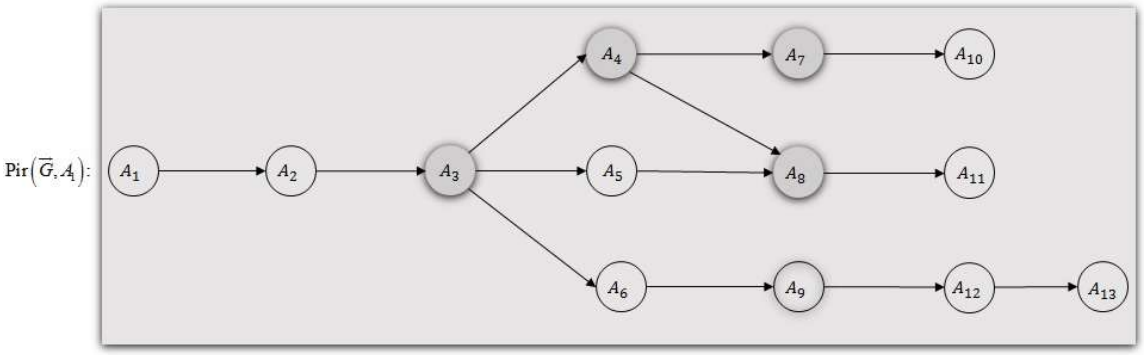

Figure 1. Technological pyramid $\mathrm{Pir}\left(\vec{G},A_1\right)$ with a root vertex A_1

The fork-vertices in the represented technological pyramid are A_3, A_4. That is, with the reaching the aggregates A_3, A_4 TR can be changed in order to improve the quality of the expected unit of production. The set of TMs contained in the technological pyramid $\mathrm{Pir}\left(\vec{G},A_1\right)$, corresponds to a set of directed paths:

$$P_1=\left(A_1,A_2,A_3,A_4,A_7,A_{10}\right),$$

$$P_2=\left(A_1,A_2,A_3,A_4,A_8,A_{11}\right),$$

$$P_3=\left(A_1,A_2,A_3,A_5,A_8,A_{11}\right),$$

$$P_4 = \left(A_1, A_2, A_3, A_6, A_9, A_{12}, A_{13}\right).$$

Each CTR in the pyramid $\mathrm{Pir}\left(\vec{G}, A_1\right)$ follows a certain TR, and before reaching the fork-vertex some CTRs should follow the same TR. In other words, the CTR is a sequence of aggregates given together with the parameters for a unit of output for each vertex in the directed path. It is clear that for each CTR the parameters of the unit of production will differ even for the same aggregate. Thus, at each time-moment, a number of CTRs will be generated for each TR in the infrastructure graph. For the technological pyramid $\mathrm{Pir}\left(\vec{G}, A_1\right)$ under consideration, we will have the following types of productive CTRs:

$$AI_1 = \left(A_1, \mathrm{P}_{11}, A_2, \mathrm{P}_{12}, A_3, \mathrm{P}_{13}, A_4, \mathrm{P}_{14}, A_7, \mathrm{P}_{17}, A_{10}, \mathrm{P}_{110}\right),$$

$$AI_2 = \left(A_1, \mathrm{P}_{21}, A_2, \mathrm{P}_{22}, A_3, \mathrm{P}_{23} A_4, \mathrm{P}_{24}, A_8, \mathrm{P}_{28}, A_{12}, \mathrm{P}_{211}\right),$$

etc.

A set of productive CTRs are formed continuously in the stream mode at each control time-moment. At the same time, the number of parameters for each productive CTR can reach a dimension of several tens or even hundreds, and the number of CTR in the training sample can achieve several hundred and tens of thousands and more. Thus, the solution of the problem of predicting the expected quality of the product (the choosing of the TR for continuation of production) requires researches in the field of processing and analyzing of Big Data.

Let at some time-moment t a set of productive CTRs $\mathrm{P}_{\mathrm{CTR}}\left(A_0, t\right) = \{AI_i : i = [1, q]\}$ with an initial vertex A_0 be formed such that their sections up to a certain fork-vertex A' coincide in the part of the passage of the aggregates. We divide the set of all CTRs $\{AI_i : i = [1, q]\}$ into several classes

$$\mathrm{P}_{\mathrm{CTR}}\left(A_0, t\right) = \mathrm{P}^{(1)}\left(A_0, t\right) \dot{\cup} \mathrm{P}^{(2)}\left(A_0, t\right) \dot{\cup} \mathrm{P}^{(3)}\left(A_0, t\right) \dot{\cup} \ldots \dot{\cup} \mathrm{P}^{(k)}\left(A_0, t\right) \tag{1}$$

such that AI_i and AI_j belong to the same class if and only if $P_i = P_j$, that is, the productive CTRs belong to the same class, if they follow the same TR (they coincide in the part of the passage of the aggregates during the whole period of the route). In the example under consideration, we have a partition:

$$\mathrm{P}_{\mathrm{CTR}}\left(A_1, t\right) = \mathrm{P}^{(1)}\left(A_1, t\right) = \left\{AI_1^i\right\} \dot{\cup} \mathrm{P}^{(2)}\left(A_1, t\right) = \left\{AI_2^i\right\} \dot{\cup}$$

$$\dot{\cup}\, \mathrm{P}^{(3)}\left(A_1, t\right) = \left\{AI_3^i\right\} \dot{\cup} \mathrm{P}^{(4)}\left(A_1, t\right) = \left\{AI_4^i\right\},$$

while the productive CTRs AI^i and AI^j belonging to the same class differ only in the part of the parameter sets for each aggregate in the sequence. It is clear that for each class in the partition (1) there is a TR common to all productive CTRs in the class. Due to the fact that the TR which belongs to the pyramid $\mathrm{Pir}\left(\vec{G}, A_1\right)$, has a common section $\left(A_1, A_2, A_3\right)$ up to the fork-vertex A_3, an applied problem of choosing a technological route arises in the process of its implementation. The substantive meaning of the problem is that for a specific TR implementation according to the initial problem it may turn out that further implementation of this TR may prove to be economically impractical, since the expected quality of the received PU will not be high enough at the planned costs for the implementation of this TR. Thus, the problem of predicting the quality of products

of metallurgical production is to determine which of the technological routes $P_i : i = [1,k]$, should be chosen for further follow-up when the fork-vertex A' is reached. To develop effective algorithms for choosing the optimal route for the continuation of production, we reduce the study to the problem of pattern recognition in a geometric formulation.

Let us consider the formulation of the problem. Consider the partition (1) of the CTR in the technological pyramid $\mathrm{Pir}\left(\vec{G}, A_0\right)$. For each class of productive CTRs $\mathrm{P}^{(i)}\left(A_0, t\right), i = [1,k]$, in the partition (1) we compile a sample

$$Z\left(P_i\right): \left(Ef\left(AI_j\right) = \frac{\mathrm{Price}\left(term\left(AI_j\right)\right) - \mathrm{C}\left(term\left(AI_j\right)\right)}{\mathrm{C}\left(term\left(AI_j\right)\right)}, i, BI_j \right), \tag{2}$$

where BI_j is the CTR on the section up to the fork-vertex A' and $Ef\left(AI_j\right)$ is defined in Definition 2.7. If the set of values of efficiency $Ef\left(AI_j\right)$ is divided into several discrete intervals $E_1, E_2, \ldots, E_m$ then the prediction problem consists in determining into which class $E_1, \ldots, E_m$ the PU falls when the TR is continued in the class $i, i = [1, \mathrm{m}]$.

Each productive CTR in the sample (2) can be represented as a multidimensional vector

$$\mathbf{a}_i = \left(a_{i_0}, a_{i_1}, a_{j_1}, \ldots, a_{j_k}\right),$$

where $a_{i_0} \in \{E_1, \ldots, E_m\}$ is the value of the efficiency of the CTR, a_{i_1} is the TR identifier of this CTR, and the values of the variables $a_{j_1}, \ldots, a_{j_k}$ correspond to the aggregates and parameters of the PU in the CTR BI_i. Thus, for a given set of n - dimensional vectors, divided into m classes,

$$\mathrm{A} = \left\{\left(a_{i_1}, a_{i_2}, \ldots, a_{i_n}\right)\right\} = \mathrm{A}_1 \cup \mathrm{A}_2 \cup \ldots \cup \mathrm{A}_m ,$$

it is required to construct a decision rule for assigning an arbitrary input vector $\mathbf{b}_i$ to one of the classes $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_m$.

Suppose that for each class of productive CTRs $\mathrm{P}^{(i)}\left(A', t\right), i = [1,k]$, a representative sample of historical CTRs has been compiled, for which the considered problem of forecasting the quality of products is solved while continuing the technological process in the i-th variant. Then the method of practical use of the proposed method is as follows. The current TR is executed before reaching the fork-vertex A'. As a result, we get CTR BI_s on the section $\left(A_0, A_1, \ldots, A'\right)$. Then, by consistently applying the found decision rules for each of the k classes for the CTR BI_s, we obtain sequences of values of the predicted value Ef for all variants of the continuation of the technological process after passing the fork-vertex A'. The obtained values Ef can be effectively applied for choosing the variant of the continuation of the technological process. For example, we can choose the variant of continuing the technological process for some Ef which takes the maximum value. More complex variants of the continuation of the technological process, can be constructed if we take into account the fulfillment of the planned indicators for the volumes of production of various types of products.

4. The algorithm for optimal partitioning of the set of classes for constructing the decision tree

In [1], a special case of the problem of predicting the quality of production is investigated. The set of values of the efficiency of productive CTRs (quality of the final PU) is divided into 2 classes of fit and defective parts. This formulation was named in the paper as the problem of predicting the defectiveness of the production where it is required to determine whether the final PU is fit if the current TR continues. The proposed approach is based on reducing the problem of pattern recognition in a geometric formulation to the research of a system of linear inequalities.

The more general formulation of the problem is considered in [9]. In this paper, an algorithm for constructing a decision function for the node of the decision tree $G=(V,E)$ is presented. The main idea of the algorithm is to achieve the maximum value of the separating function. The training set of vectors would be separated by a hyperplane

$$f(\mathbf{a}) = \mathbf{a}\cdot n_v - \varepsilon_v = 0,$$

where v is the vertex of the decision tree with the given sample A_v, and the vectors n_v, ε_v are given in such a way that the descendants v_1, v_2 of the vertex v in the decision tree correspond to the samples

$$A_{v_1} = \left\{\mathbf{a}_j \in A_v : \langle n_v, \mathbf{a}_j\rangle \text{,, } \varepsilon_v\right\} \neq \varnothing,$$

$$A_{v_2} = \left\{\mathbf{a}_j \in A_v : \langle n_v, \mathbf{a}_j\rangle > \varepsilon_v\right\} \neq \varnothing.$$

Reaching maximum of the value

$$\operatorname{discrim}(A_v, n_v, \varepsilon_v) = \sum_{i\in I}\left|\frac{\left|A_{v_1}\cap A_i\right|}{\left|A_{v_1}\right|} - \frac{\left|A_{v_2}\cap A_i\right|}{\left|A_{v_2}\right|}\right|$$

provides an effective partition of the set of vectors according to the classes of the training sample. Thus, the practical implementation of the approach requires the choice of partition A_v and vectors n_v, ε_v such that the value $\operatorname{discrim}(A_v, n_v, \varepsilon_v)$ reaches its maximum.

In this paper, we present an algorithm for partitioning a sample A_v into two classes $I = I_1 \dot\cup I_2$, an important advantage of which is a significant reduction in the amount of computation when choosing the optimal partition. The result of the implementation of the developed algorithm serves as a basis for constructing a decision tree, while the choice of vectors n_v, ε_v for assigning an arbitrary vector to a certain class of training sample can be carried out using the algorithm proposed in [9].

Let us assume that a sample $\mathrm{A} = \mathrm{A}_1 \dot\cup \mathrm{A}_2 \dot\cup \ldots \dot\cup \mathrm{A}_m$ is given, a set of classes $I = [1,m]$ and $I_i = \left\{j : \mathbf{a}_j \in \mathrm{A}_i\right\}$. We introduce the notation:

$$a_{ij} = \sum\left\{\left(\mathbf{a}_{k_1} - \mathbf{a}_{k_2}\right) : \mathbf{a}_{k_1} \in \mathrm{A}_i, \mathbf{a}_{k_2} \in \mathrm{A}_j\right\}, i, j \in [1,k].$$
$$a(I_1, I_2) = \sum_{i\in I_1, j\in I_2} a_{ij}. \tag{3}$$

Among all partitions of the form

$$I_1, I_2 \subseteq [1,m] : I_1, I_2 \neq \varnothing,\ I_1 \dot\cup I_2 = I,$$

for the sample A_v given for the vertex v of the decision tree $G=(V,E)$, it is required to choose such one that the value $\|a(I_1,I_2)\|$ reaches its maximum, that is

$$\|a(I_1,I_2)\| \to \max . \tag{4}$$

Suppose that for a training sample $A = A_1 \dot\cup \ldots \dot\cup A_m$ the sum of differences of the form (3) is calculated pairwise for all classes. To search for a partition $I = I_1 \dot\cup I_2$, which provides a maximum for the function $a(I_1,I_2)$, we would consider the following heuristic algorithm.

Input data: the set of classes $I=[1,m]$

Output data: the partition $I = I' \dot\cup I''$

Let $I' = \varnothing, I'' = I$

While $I'' \neq \varnothing$ **do**

$i_0 = \arg\max\limits_{i \in I''} \left(\left\{ \| a(I' \dot\cup \{i\}, I'', \{i\}) \| \right\} \right)$

$I' \leftarrow I' \dot\cup \{i_0\}$

$I'' \leftarrow I'', \{i_0\}$

End of condition

End of algorithm

In the process of the algorithm, it is constructed a sequence of partitions of the form

$$I_1 = \varnothing \dot\cup I ,$$

$$I_2 = \{i_1\} \dot\cup I , \ \{i_1\} ,$$

$$\ldots\ldots\ldots\ldots ,$$

$$I_n = I \dot\cup \varnothing ,$$

among which one should choose the such one that $\|a(I',I'')\|$ is the maximal. Thus, for a given training sample and a set of classes, a partition is constructed taking into account the maximization of the value $\|a(I_1,I_2)\|$. In this case, the partition is the basis for constructing a decision tree for classifying vectors from the training sample.

Let's consider an example where the search for a solution is carried out using the developed approach.

Example 4.1 Let there be given a set $I=\{1,2,3,4\}$ and 4 vectors a_1,a_2,a_3,a_4.

Then the matrix of values for a_{ij} will have the form

$$a_{ij} = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ 1 & 0 & (1,0) & (0,-1) & (1,-1) \\ 2 & (0,-1) & 0 & (-1,-1) & (0,-1) \\ 3 & (0,1) & (1,1) & 0 & (1,0) \\ 4 & (-1,1) & (0,1) & (-1,0) & 0 \end{array} .$$

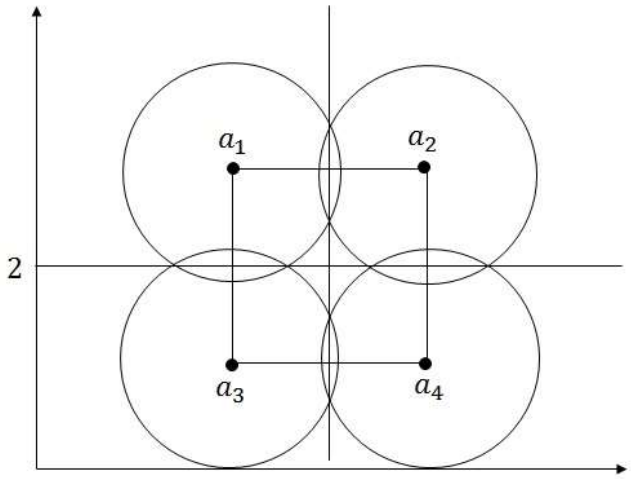

Figure 2. A partitioning the set of vectors into classes

Using the developed approach, we construct the optimal partition of vectors into classes.

1. $I_1 = \{1\}, I_2 = \{2,3,4\}$,
   $a(I_1, I_2) = a_{12} + a_{13} + a_{14} = (1,0) + (0,-1) + (1,-1) = (2,-2)$,
   $\| a(I_1, I_2) \| = \sqrt{4+4} = 2\sqrt{2} \quad 2,8$.
2. $I_1 = \{1,2\}, I_2 = \{3,4\}$,
   $a(I_1, I_2) = a_{13} + a_{14} + a_{23} + a_{24} = (0,-1) + (1,-1) + (-1,-1) + (0,-1) = (0,-4)$,
   $\| a(I_1, I_2) \| = 4$.
3. $I_1 = \{1,2,3\}, I_2 = \{4\}$,
   $a(I_1, I_2) = a_{14} + a_{24} + a_{34} = (1,-1) + (0,-1) + (1,0) = (2,-2)$,
   $\| a(I_1, I_2) \| = \sqrt{4+4} = 2\sqrt{2} \quad 2,8$.

The maximal value $a(I_1, I_2)$ is achieved for the partition 2, that is

$$I = \{1,2\} \cup \{3,4\},$$

which looks like a logical result in a practical sense.

The proposed search algorithm for optimal partitioning is effective due to the use of a matrix $\| a_{ij} \|_1^k$ with previously calculated values a_{ij}. When processing huge amount of historical data, where the sample size can achieve hundreds of thousands and millions of vectors, it is advisable to apply the method proposed in the present work not to the initial sample, but to a sample obtained from the initial one as a result of sample clustering, for example, using the effective VNS-clustering method from [10] with the replacement of each cluster by the n - dimensional vector representing it.

5. Conclusion

The paper considers the applied problem of optimization of technological processes in metallurgical production. Historical data generated by the implementation of productive CTRs forms the technological database, and, within the framework of the developed approach, a training sample of the supervised learning. An approach is proposed to solve the problem of clustering the set of training sample vectors, which allows to significantly narrow the solution search space and reduce

computational costs. The main idea of the approach is the introduction of a special criterion for the optimality of the partition, on the basis of which an algorithm for search a solution close to the optimal has been developed. The algorithm is implemented in a test case, and the results obtained are the basis for constructing a decision tree for the initial classification problem.

References

[1] Gainanov D N and Berenov D A 2017 Algorithm for predicting the Quality of the product of Metallurgical Production *Proc. Int. Conf. on Big Data and Advanced Analytics* (Minsk) p 65-70
[2] Gainanov D N 2014 *Combinatorial Geometry and Graphs in the Analysis of Infeasible Systems and Pattern Recognition* (Moscow: Nauka)
[3] Gainanov Damir N 2016 Graphs for Pattern Recognition. *Infeasible Systems of Linear Inequalities* (DeGruyter)
[4] Gainanov D N 1985 On combinatorial properties of infeasible systems of linear inequalities and convex polyhedra *Math notices* vol 38 3 p 463–474
[5] Mazurov Vl D 1990 *Committees method in problems of optimization and classification* (Moscow: Nauka)
[6] Mazurov Vl D and Khachai M Yu 2004 Committees of systems of linear inequalities *Automation and Remote Control* 2 p 43–54
[7] Gainanov D N 1992 *Alternative Covers and Independence Systems in Pattern Recognition Pattern Recognition and Image Analysis* vol 2 **2** p 147–160
[8] Gainanov D N and Matveev A O 1991 Lattice Diagonals and Geometric Pattern Recognition Problems *Pattern Recognition and Image Analysis* vol 1 **3** p 277–282
[9] Gainanov D N and Berenov D A 2017 *Algorithm for Predicting the Quality of the Product of Metallurgical Production Proc. Int. Conf. on Optimization and Applications* (Pertovac) accepted for publication
[10] Hansen P and Mladenovic N 2001 J-MEANS: a new local search heuristic for minimum sum of squares clustering *Pattern Recognition* **34** p 405–413